ARTICLE 29 DATA PROTECTION WORKING PARTY

0829/14/EN
WP216

Opinion 05/2014 on Anonymisation Techniques

Adopted on 10 April 2014

THE WORKING PARTY ON THE PROTECTION OF INDIVIDUALS WITH REGARD TO THE PROCESSING OF PERSONAL DATA

set up by Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995,

having regard to Articles 29 and 30 thereof,

having regard to its Rules of Procedure,

HAS ADOPTED THE PRESENT OPINION:

EXECUTIVE SUMMARY

In this Opinion, the WP analyses the effectiveness and limits of existing anonymisation techniques against the EU legal background of data protection and provides recommendations to handle these techniques by taking account of the residual risk of identification inherent in each of them.

The WP acknowledges the potential value of anonymisation in particular as a strategy to reap the benefits of 'open data' for individuals and society at large whilst mitigating the risks for the individuals concerned. However, case studies and research publications have shown how difficult it is to create a truly anonymous dataset whilst retaining as much of the underlying information as required for the task.

In the light of Directive 95/46/EC and other relevant EU legal instruments, anonymisation results from processing personal data in order to irreversibly prevent identification. In doing so, several elements should be taken into account by data controllers, having regard to all the means "likely reasonably" to be used for identification (either by the controller or by any third party).

Anonymisation constitutes a further processing of personal data; as such, it must satisfy the requirement of compatibility by having regard to the legal grounds and circumstances of the further processing. Additionally, anonymized data do fall out of the scope of data protection legislation, but data subjects may still be entitled to protection under other provisions (such as those protecting confidentiality of communications).

The main anonymisation techniques, namely randomization and generalization, are described in this opinion. In particular, the opinion discusses noise addition, permutation, differential privacy, aggregation, k-anonymity, l-diversity and t-closeness. It explains their principles, their strengths and weaknesses, as well as the common mistakes and failures related to the use of each technique.

The opinion elaborates on the robustness of each technique based on three criteria:

(i) is it still possible to single out an individual,

(ii) is it still possible to link records relating to an individual, and

(iii) can information be inferred concerning an individual?

Knowing the main strengths and weaknesses of each technique helps to choose how to design an adequate anonymisation process in a given context.

Pseudonymisation is also addressed to clarify some pitfalls and misconceptions: pseudonymisation is not a method of anonymisation. It merely reduces the linkability of a dataset with the original identity of a data subject, and is accordingly a useful security measure.

The Opinion concludes that anonymisation techniques can provide privacy guarantees and may be used to generate efficient anonymisation processes, but only if their application is engineered appropriately – which means that the prerequisites (context) and the objective(s) of the anonymisation process must be clearly set out in order to achieve the targeted anonymisation while producing some useful data. The optimal solution should be decided on

a case-by-case basis, possibly by using a combination of different techniques, while taking into account the practical recommendations developed in this Opinion.

Finally, data controllers should consider that an anonymised dataset can still present residual risks to data subjects. Indeed, on the one hand, anonymisation and re-identification are active fields of research and new discoveries are regularly published, and on the other hand even anonymised data, like statistics, may be used to enrich existing profiles of individuals, thus creating new data protection issues. Thus, anonymisation should not be regarded as a one-off exercise and the attending risks should be reassessed regularly by data controllers.

1 Introduction

While devices, sensors and networks create large volumes and new types of data, and the cost of data storage is becoming negligible, there is a growing public interest in and demand for the re-use of these data. 'Open data' may provide clear benefits for society, individuals and organisations, but only if everybody's rights are respected to the protection of their personal data and private life.

Anonymisation may be a good strategy to keep the benefits and to mitigate the risks. Once a dataset is truly anonymised and individuals are no longer identifiable, European data protection law no longer applies. However, it is clear from case studies and research publications that the creation of a truly anonymous dataset from a rich set of personal data, whilst retaining as much of the underlying information as required for the task, is not a simple proposition. For example, a dataset considered to be anonymous may be combined with another dataset in such a way that one or more individuals can be identified.

In this Opinion, the WP analyses the effectiveness and limits of existing anonymisation techniques against the EU legal background of data protection and provides recommendations for a cautious and responsible use of these techniques to build a process of anonymisation.

2 Definitions & Legal Analysis

2.1. Definitions in the EU Legal Context

Directive 95/46/EC refers to anonymisation in Recital 26 to exclude anonymised data from the scope of data protection legislation:

> *"Whereas the principles of protection must apply to any information concerning an identified or identifiable person; whereas, to determine whether a person is identifiable, account should be taken of all the means likely reasonably to be used either by the controller or by any other person to identify the said person; whereas the principles of protection shall not apply to data rendered anonymous in such a way that the data subject is no longer identifiable; whereas codes of conduct within the meaning of Article 27 may be a useful instrument for providing guidance as to the ways in which data may be rendered anonymous and retained in a form in which identification of the data subject is no longer possible;".*[1]

Close reading of Recital 26 provides a conceptual definition of anonymisation. Recital 26 signifies that to anonymise any data, the data must be stripped of sufficient elements such that the data subject can no longer be identified. More precisely, thet data must be processed in such a way that it can no longer be used to identify a natural person by using "all the means likely reasonably to be used" by either the controller or a third party. An important factor is that the processing must be irreversible. The Directive does not clarify how such a de-identification process should or could be performed[2]. The focus is on the outcome: that data should be such as not to allow the data subject to be identified via "all" "likely" and "reasonable" means. Reference is made to codes of conduct as a tool to set out possible

[1] It should be noted, in addition, that this is the approach also followed in the draft EU data protection Regulation, under Recital 23 "to determine whether a person is identifiable, account should be taken of all the means likely reasonably to be used either by the controller or by any other person to identify the individual".

[2] This concept is elaborated further on p. 8 of this Opinion.

anonymisation mechanisms as well as retention in a form in which identification of the data subject is “no longer possible”. The Directive thus clearly sets a very high standard.

The e-Privacy Directive (Directive 2002/58/EC) also refers to “anonymisation” and “anonymous data” very much in the same regard. Recital 26 states that:

> *“Traffic data used for marketing communications services or for the provision of value added services should also be erased or made anonymous after the provision of the service”.*

Accordingly, Article 6(1) states that:

> *“Traffic data relating to subscribers and users processed and stored by the provider of a public communications network or publicly available electronic communications service must be erased or made anonymous when it is no longer needed for the purpose of the transmission of a communication without prejudice to paragraphs 2, 3 and 5 of this Article and Article 15(1).”*

Under Article 9(1), moreover:

> *“Where location data other than traffic data, relating to users or subscribers of public communications networks or publicly available electronic communications services, can be processed, such data may only be processed when they are made anonymous, or with the consent of the users or subscribers to the extent and for the duration necessary for the provision of a value added service.”*

The underlying rationale is that the outcome of anonymisation as a technique applied to personal data should be, in the current state of technology, as permanent as erasure, i.e. making it impossible to process personal data.[3]

2.2. Legal Analysis

Analysis of the wording related to anonymisation in the leading EU data protection instruments allows highlighting four key features:

- Anonymisation can be a result of processing personal data with the aim of irreversibly preventing identification of the data subject.

- Several anonymisation techniques may be envisaged, there is no prescriptive standard in EU legislation.

- Importance should be attached to contextual elements: account must be taken of “all” the means “likely reasonably” to be used for identification by the controller and third parties, paying special attention to what has lately become, in the current state of technology, “likely reasonably” (given the increase in computational power and tools available).

[3] It should be recalled here that anonymisation is also defined in international standards such as the ISO 29100 one – being the “Process by which personally identifiable information (PII) is irreversibly altered in such a way that a PII principal can no longer be identified directly or indirectly, either by the PII controller alone or in collaboration with any other party” (ISO 29100:2011). Irreversibility of the alteration undergone by personal data to enable direct or indirect identification is the key also for ISO. From this standpoint, there is considerable convergence with the principles and concepts underlying the 95/46 Directive. This also applies to the definitions to be found in some national laws (for instance, in Italy, Germany and Slovenia), where the focus is on non-identifiability and reference is made to the “disproportionate effort” to re-identify (D, SI). However, the French Data Protection Law provides that data remains personal data even if it is extremely hard and unlikely to re-identify the data subject – that is to say, there is no provision referring to the “reasonableness” test.

- A risk factor is inherent to anonymisation: this risk factor is to be considered in assessing the validity of any anonymisation technique - including the possible uses of any data that is "anonymised" by way of such technique - and severity and likelihood of this risk should be assessed.

In this opinion the notation "anonymisation technique" is used, rather than "anonymity" or "anonymous data", to point out the inherent residual risk of re-identification linked to any technical-organizational measure aimed at rendering data "anonymous".

2.2.1. Lawfulness of the Anonymisation Process

First, anonymisation is a technique applied to personal data in order to achieve irreversible de-identification. Therefore, the starting assumption is that the personal data must have been collected and processed in compliance with the applicable legislation on the retention of data in an identifiable format.

In this context, the anonymisation process, meaning the processing of such personal data to achieve their anonymisation, is an instance of "further processing". As such, this processing must comply with the test of compatibility in accordance with the guidelines provided by the Working Party in its Opinion 03/2013 on purpose limitation[4].

This means that, in principle, the legal basis for anonymisation can be found in any of the grounds mentioned in Article 7 (including the data controller's legitimate interest) provided the data quality requirements of Article 6 of the Directive are also met and with due regard to the specific circumstances and all the factors mentioned in the Working Party's opinion on purpose limitation[5].

On the other hand, the provisions contained in Article 6(1) e) of the Directive 95/46/EC (but also in Articles 6(1) and 9(1) of the e-Privacy Directive) ought to be pointed out as they demonstrate the need to keep personal data "in a form which permits identification" for no longer than is necessary for the purposes of the collection or further processing.

In itself, this provision makes a strong point that personal data should, at least, be anonymised "by default" (subject to different legal requirements, such as those mentioned in the e-Privacy Directive regarding traffic data). If the data controller wishes to retain such personal data once the purposes of the original or further processing have been achieved, anonymisation techniques should be used so as to irreversibly prevent identification.

Accordingly, the Working Party considers that anonymisation as an instance of further processing of personal data can be considered to be compatible with the original purposes of the processing but only on condition the anonymisation process is such as to reliably produce anonymised information in the sense described in this paper.

---

[4] Opinion 03/2013 of the Article 29 Working Party, available at: http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf

[5] This means, in particular, that a substantive assessment has to be carried out in the light of all relevant circumstances with particular regard to the following key factors:
a) the relationship between the purposes for which the personal data have been collected and the purposes of further processing;
b) the context in which the personal data have been collected and the reasonable expectations of the data subjects as to their further use;
c) the nature of the personal data and the impact of the further processing on the data subjects;
d) the safeguards adopted by the controller to ensure fair processing and to prevent any undue impact on the data subjects.

It should also be emphasized that anomymisation has to be held in compliance with the legal constraints recalled by the European Court of Justice in its decision on case C-553/07 (*College van burgemeester en wethouders van Rotterdam v M.E.E. Rijkeboer*), relating to the need to retain the data in an identifiable format to enable, for instance, the exercise of access rights by data subjects. The ECJ ruled that "*Article 12(a) of the [95/46] Directive requires Member States to ensure a right of access to information on the recipients or categories of recipient of personal data and on the content of the data disclosed not only in respect of the present but also in respect of the past. It is for Member States to fix a time-limit for storage of that information and to provide for access to that information which constitutes a fair balance between, on the one hand, the interest of the data subject in protecting his privacy, in particular by way of his rights to object and to bring legal proceedings and, on the other, the burden which the obligation to store that information represents for the controller*."

This is especially relevant in case Article 7(f) of the 95/46 Directive is relied upon by a data controller as regards anonymisation: the data controller's legitimate interest must be always balanced against the data subjects' rights and fundamental freedoms.

For instance, an investigation by the Dutch DPA in 2012-2013 into the use of deep packet inspection technologies by four mobile operators showed a legal ground under Article 7(f) of Directive 95/46 for the anonymisation of the contents of traffic data as soon as possible after the collection of these data. Indeed, Article 6 of the e-Privacy Directive stipulates that traffic data relating to subscribers and users processed and stored by the provider of a public communications network or publicly available electronic communications service must be erased or made anonymous as soon as possible. In this case, because it is allowed under Article 6 of the e-Privacy Directive, a corresponding legal ground in Article 7 of the Data Protection Directive exists. This could also be presented the other way around: if a type of data processing is not allowed under Article 6 of the e-Privacy Directive, there cannot be a legal ground in Article 7 of the Data Protection Directive.

2.2.2. Potential Identifiability of Anonymised Data

The Working Party has addressed the concept of personal data in detail in Opinion 4/2007 on personal data, focusing on the building blocks of the definition contained in Article 2(a) of Directive 95/46/EC including the "identified or identifiable" part of such definition. In this context, the Working Party has also concluded that "Anonymised data would therefore be anonymous data that previously referred to an identifiable person, but where that identification is no longer possible".

The Working Party has therefore already clarified that the "means … reasonably to be used" test is suggested by the Directive as a criterion to be applied in order to assess whether the anonymisation process is sufficiently robust, i.e. whether identification has become "reasonably" impossible. The particular context and circumstances of a specific case directly impact on identifiability. In the Technical Annex to this opinion, analysis is provided on the impact of choosing the most appropriate technique.

As already emphasized, research, tools and computational power evolve. Therefore, it is neither possible nor useful to provide an exhaustive enumeration of circumstances when identification is no longer possible. However, some key factors deserve to be taken into consideration and illustrated.

Firstly, it can be argued that data controllers should focus on the concrete means that would be necessary to reverse the anonymisation technique, notably regarding the cost and the

know-how needed to implement those means and the assessment of their likelihood and severity. For instance, they should balance their anonymisation effort and costs (in terms of both time and resources required) against the increasing low-cost availability of technical means to identify individuals in datasets, the increasing public availability of other datasets (such as those made available in connection with 'Open data' policies), and the many examples of incomplete anonymisation entailing subsequent adverse, sometimes irreparable effects on data subjects.[6] It should be noted that the identification risk may increase over time and depends also on the development of information and communication technology. Legal regulations, if any, must therefore be formulated in a technologically neutral manner and ideally take into account the changes in the developing potentials of information technology.[7]

Secondly, “the means likely reasonably to be used to determine whether a person is identifiable” are those to be used “by the controller or by any other person”. Thus, it is critical to understand that when a data controller does not delete the original (identifiable) data at event-level, and the data controller hands over part of this dataset (for example after removal or masking of identifiable data), the resulting dataset is still personal data. Only if the data controller would aggregate the data to a level where the individual events are no longer identifiable, the resulting dataset can be qualified as anonymous. For example: if an organisation collects data on individual travel movements, the individual travel patterns at event level would still qualify as personal data for any party, as long as the data controller (or any other party) still has access to the original raw data, even if direct identifiers have been removed from the set provided to third parties. But if the data controller would delete the raw data, and only provide aggregate statistics to third parties on a high level, such as 'on Mondays on trajectory X there are 160% more passengers than on Tuesdays', that would qualify as anonymous data.

An effective anonymisation solution prevents all parties from singling out an individual in a dataset, from linking two records within a dataset (or between two separate datasets) and from inferring any information in such dataset. Generally speaking, therefore, removing directly identifying elements in itself is not enough to ensure that identification of the data subject is no longer possible. It will often be necessary to take additional measures to prevent identification, once again depending on the context and purposes of the processing for which the anonymised data are intended.

---

[6] Interestingly, the European Parliament’s amendments to the Draft General Data Protection Regulation as submitted recently (21 October 2013) mention specifically in Recital 23 that “To ascertain whether means are reasonably likely to be used to identify the individual, account should be taken of all objective factors, <u>such as the costs of and the amount of time required for identification, taking into consideration both available technology at the time of the processing and technological development</u>”.

[7] See Opinion 4/2007 of the Article 29 Working Party, p. 15.

EXAMPLE:
Genetic data profiles are an example of personal data that can be at risk of identification if the sole technique used is the removal of the identity of the donor due to the unique nature of certain profiles. It has already been shown in the literature[8] that the combination of publically available genetic resources (e.g. genealogy registers, obituary, results of search engine queries) and the metadata about DNA donors (time of donation, age, place of residence) can reveal the identity of certain individuals even if that DNA was donated "anonymously".

Both families of anonymisation techniques – data randomization and generalization –[9] have shortcomings; however, each of them may be appropriate under the given circumstances and context to achieve the desired purpose without jeopardising data subjects' privacy. It must be clear that 'identification' not only means the possibility of retrieving a person's name and/or address, but also includes potential identifiability by singling out, linkability and inference. Furthermore, for data protection law to apply, it does not matter what the intentions are of the data controller or recipient. As long as the data are identifiable, data protection rules apply.

Where a third party processes a dataset treated with an anonymisation technique (anonymised and released by the original data controller) they may do so lawfully without needing to take account of data protection requirements provided they cannot (directly or indirectly) identify the data subjects in the original dataset. However, third parties are required to take account of any contextual and circumstantial factors mentioned above (including the specific features of the anonymisation techniques as applied by the original data controller) in deciding how to use and, especially, combine such anonymised data for their own purposes – as the resulting consequences may entail different types of liability on their part. Where those factors and features are such as to entail an unacceptable risk of identification of data subjects, the processing will fall once again under the scope of data protection law.

The above list is in no way meant to be exhaustive, but rather to provide general guidance on the approach to assessing the identifiability potential of a given dataset that undergoes anonymisation according to the different available techniques. All of the above factors can be considered to be as many risk factors to be weighed both by data controllers in anonymising datasets and by third parties in using those "anonymised" datasets for their own purposes.

2.2.3. Risks of the Use of Anonymised Data

When considering using anonymisation techniques, data controllers have to take into account the following risks

- A specific pitfall is to consider pseudonymised data to be equivalent to anonymised data. The Technical Analysis section will explain that pseudonymised data cannot be equated to anonymised information as they continue to allow an individual data subject to be singled out and linkable across different data sets. Pseudonymity is likely to allow for identifiability, and therefore stays inside the scope of the legal regime of data protection. This is especially relevant in the context of scientific, statistical or historical research.[10]

[8] See John Bohannon, Genealogy Databases Enable Naming of Anonymous DNA Donors, Science, Vol. 339, No. 6117 (18 January 2013), p. 262.
[9] The main features and differences of these two anonymisation techniques are described in Section 3 below ("Technical Analysis").
[10] See also Opinion 4/2007 of the Article 29 Working Party, p. 18-20.

EXAMPLE:
A typical instance of the misconceptions surrounding pseudonymisation is provided by the well-known "AOL (America On Line) incident". In 2006, a database containing twenty million search keywords for over 650,000 users over a 3-month period was publically released, with the only privacy preserving measure consisting in replacing AOL user ID by a numerical attribute. This led to the public identification and location of some of them. Pseudonymised search engine query strings, especially if coupled with other attributes, such as IP addresses or other client configuration parameters, possess a very high power of identification.

- A second mistake is to consider that properly anonymised data (having satisfied all the conditions and criteria mentioned above and falling by definition outside the scope of the data protection Directive) deprive individuals of whatever safeguards – first and foremost, because other pieces of legislation may apply to the use of these data. For example, Article 5(3) of the e-Privacy Directive prevents storage of and access to "information" of any type (including non-personal information) on terminal equipment without the subscriber's/user's consent as this is part of the broader principle of confidentiality of communications.

- A third negligence would also result from not considering the impact on individuals, under certain circumstances, by properly anonymised data, especially in the case of profiling. The sphere of an individual's private life is protected by Article 8 of the ECHR and Article 7 of the EU Charter of Fundamental Rights; as such, even though data protection laws may no longer apply to this type of data, the use made of datasets anonymised and released for use by third parties may give rise to a loss of privacy. Special caution is required in handling anonymised information especially whenever such information is used (often in combination with other data) for taking decisions that produce effects (albeit indirectly) on individuals. As already pointed out in this Opinion and clarified by the Working Party in particular in the Opinion on the concept of "purpose limitation" (Opinion 03/2013)[11], the data subjects' legitimate expectations concerning further processing of their data should be assessed in the light of the relevant context-related factors – such as the nature of the relationship between data subjects and data controllers, applicable legal obligations, transparency of processing operations.

3 Technical Analysis, Robustness of Technologies and Typical Mistakes

Different anonymisation practices and techniques exist with variable degrees of robustness. This section will address the main points to be considered by data controllers in applying them by having regard, in particular, to the guarantee attainable by the given technique taking into account the current state of technology and considering three risks which are essential to anonymisation:

- *Singling out* , which corresponds to the possibility to isolate some or all records which identify an individual in the dataset;
- *Linkability*, which is the ability to link, at least, two records concerning the same data subject or a group of data subjects (either in the same database or in two different databases). If an attacker can establish (e.g. by means of correlation analysis) that two records are assigned to a same group of individuals but cannot single out individuals in this group, the technique provides resistance against "singling out" but not against linkability;

[11] Available at http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf

- *Inference*, which is the possibility to deduce, with significant probability, the value of an attribute from the values of a set of other attributes.

Thus, a solution against these three risks would be robust against re-identification performed by the most likely and reasonable means the data controller and any third party may employ. The Working Party emphasizes, in this connection, that techniques of de-identification and anonymisation are the subject of ongoing research and such research has shown consistently that no technique is devoid of shortcomings per se. Broadly speaking there are two different approaches to anonymisation: the first is based on ***randomization*** while the second is based on ***generalization***. The opinion also addresses other concepts like *pseudonymisation*, *differential-privacy*, *l-diversity*, *t-closeness*.

This opinion uses the following vocabulary in this section: a dataset is composed of different records relating to individuals (the data subjects). Each record is related to one data subject and is composed of a set of <u>values</u> (or “entries”, e.g.: 2013) for each <u>attribute</u> (e.g. year). A dataset is a collection of records that can be shaped alternatively as a table (or a set of tables) or as an annotated/weighted graph, which is increasingly the case today. The examples in the opinion will relate to tables, but they are applicable also to other graphical representations of records. Combinations of attributes relating to a data subject or a group of data subjects may be referred to as <u>quasi-identifiers</u>. In some cases, a dataset can have multiple records on the same individual. An “<u>attacker</u>” is a third party (i.e., neither the data controller nor the data processor) accessing the original records whether accidentally or intentionally.

3.1. Randomization

Randomization is a family of techniques that alters the veracity of the data in order to remove the strong link between the data and the individual. If the data are sufficiently uncertain then they can no longer be referred to a specific individual. Randomization by itself will not reduce the singularity of each record as each record will still be derived from a single data subject but may protect against inference attacks/risks. and can be combined with generalization techniques to provide stronger privacy guarantees. Additional techniques may be required to ensure that a record cannot identify a single individual.

3.1.1. Noise addition

The technique of noise addition is especially useful when attributes may have an important adverse effect on individuals and consists of modifying attributes in the dataset such that they are less accurate whilst retaining the overall distribution. When processing a dataset, an observer will assume that values are accurate but this will only be true to a certain degree. As an example, if an individual’s height was originally measured to the nearest centimetre the anonymised dataset may contain a height accurate to only +-10cm. If this technique is applied effectively, a third-party will not be able to identify an individual nor should he be able to repair the data or otherwise detect how the data have been modified.

Noise addition will commonly need to be combined with other anonymisation techniques such as the removal of obvious attributes and quasi-identifiers. The level of noise should depend on the necessity of the level of information required and the impact on individuals’ privacy as a result of disclosure of the protected attributes.

3.1.1.1. Guarantees

- *Singling out:* It is still possible to single out the records of an individual (perhaps in a non-identifiable manner) even though the records are less reliable.
- *Linkability:* It is still possible to link the records of the same individual, but the records are less reliable and thus a real record can be linked to an artificially added one (i.e. to 'noise'). In some cases, a wrong attribution might expose a data subject to significant and even higher level of risk than a correct one.
- *Inference:* Inference attacks may be possible but the success rate will be lower and some false positives (and false negatives) are plausible.

3.1.1.2. Common mistakes

- Adding inconsistent noise: If noise is not semantically viable (i.e. it is "out-of-scale" and does not respect the logic between attributes in a set) then an attacker having access to the database will be able to filter out the noise and, in some cases, regenerate the missing entries. Furthermore, if the dataset is too sparse[12], it can remain possible to link the noisy data entries with an external source.
- Assuming that noise addition is enough: noise addition is a complementary measure that makes it harder for an attacker to retrieve the personal data. Unless the noise is higher than the information contained in the dataset, it should not be assumed that noise addition represents a standalone solution for anonymisation.

3.1.1.3. Failures of Noise Addition

A very famous re-identification experiment is the one performed on the customers' database of the video content provider Netflix. Researchers have analyzed the geometric properties of that database consisting of more than 100 million ratings on a scale 1-5 on over 18.000 movies, expressed by almost 500.000 users, publicly released by the company, after being "anonymised" according to an internal privacy policy, with all customer identifying information removed except ratings and dates. Noise was added as the ratings were slightly increased or decreased.

In spite of this, it was found that 99% of user records could be uniquely identified in the dataset using 8 ratings and dates with 14-day errors as selection criteria, whilst lowering the selection criteria (2 ratings and 3-day error) still allowed identifying 68% of users.[13]

3.1.2. Permutation

This technique, consists of shuffling the values of attributes in a table so that some of them are artificially linked to different data subjects, is useful when it is important to retain the exact distribution of each attribute within the dataset.

Permutation may be considered as a special form of noise addition. In a classical noise technique, attributes are modified with randomized values. Generating consistent noise can be a difficult task and modifying attribute values slightly may not provide adequate privacy. As an alternative, permutation techniques alter values within in the dataset by just swapping them

[12] This concept is elaborated further in the Annex, p. 30.

[13] Narayanan, A., & Shmatikov, V. (2008, May). Robust de-anonymization of large sparse datasets. In *Security and Privacy, 2008. SP 2008. IEEE Symposium on* (pp. 111-125). IEEE.

from one record to another. Such swapping will ensure that range and distribution of values will remain the same but correlations between values and individuals will not. If two or more attributes have a logical relationship or statistical correlation and are permutated independently, such a relationship will be destroyed. It may therefore be important to permutate a set of related attributes so as to not to break the logical relationship, otherwise an attacker could identify the permutated attributes and reverse the permutation.

For instance, if we consider a subset of attributes in a medical dataset such as "reasons for hospitalization/symptoms/department in charge", a strong logical relationship will link the values in most cases and permutation of only one of the values would thus be detected and could even be reversed.

Similarly to noise addition, permutation may not provide anonymisation by itself and should always be combined with the removal of obvious attributes/quasi-identifiers.

3.1.2.1. Guarantees

- *Singling out:* As with noise addition, it is still possible to single out the records of an individual but records are less reliable.
- *Linkability:* If permutation affects attributes and quasi-identifiers, it may prevent "correct" linking of attributes both internally and externally to a dataset but still allow "incorrect" linkability, since a real entry can be associated to a different data subject.
- *Inference:* Inferences may still be drawn from the dataset, especially if attributes are correlated or have strong logical relationships; however, not knowing which attributes have been permutated, the attacker has to consider that his inference is based on a wrong hypothesis and therefore only probabilistic inference remains possible.

3.1.2.2. Common mistakes

- Selecting the wrong attribute: permutating the non-sensitive or non-risky attributes would not result in a significant gain in terms of personal data protection. Indeed if the sensitive/risky attributes were still associated to the original attribute, then an attacker would still be able to extract sensitive information about individuals.
- Permutating attributes randomly: If two attributes are strongly correlated, then permutating the attributes randomly will not provide strong guarantees. This common mistake is illustrated in Table 1.
- Assuming that permutation is enough: Like noise addition, permutation does not provide anonymity by itself and should be combined with other techniques such as the removal of obvious attributes.

3.1.2.3. Failures of permutation

This example shows how randomly permutating attributes results in poor privacy guarantees when logical links exist between different attributes. Following the attempted anonymisation, it is trivial to deduce the income of each individual depending on the job (and year of birth). For instance, it can be argued, by direct inspection of the data, that the CEO in the table was very likely born in 1957 and has the highest salary, whilst the unemployed was born in 1964 and has the lowest income.

Year	Gender	Job	Income (permuted)
1957	M	Engineer	70k
1957	M	CEO	5k
1957	M	Unemployed	43k
1964	M	Engineer	100k
1964	M	Manager	45k

Table 1. An ineffective example of anonymisation by permutation of correlated attributes

3.1.3. Differential privacy

Differential privacy[14] falls within the family of randomization techniques, with a different approach: while, in fact, noise insertion comes into play beforehand when dataset is supposed to be released, differential privacy can be used when the data controller generates anonymised views of a dataset whilst retaining a copy of the original data. Such anonymised views would typically be generated through a subset of queries for a particular third party. The subset includes some random noise deliberately added ex-post. Differential privacy tells the data controller how much noise he needs to add, and in which form, to get the necessary privacy guarantees.[15] In this context, it will be especially important to continuously monitor (at least for each new query), for any possibility to identify an individual in the query result set. It has however to be clarified that differential privacy techniques will not change the original data and thus, as long as the original data remains, the data controller is able to identify individuals in results of differential privacy queries taking into account all the means likely reasonably to be used. Such results have also to be considered as personal data.

One benefit of an approach based on differential privacy lies in the fact that datasets are provided to authorised third parties in response to a specific query rather than through the release of a single dataset. To assist with audit a list of all queries and requests can be retained by the data controller ensuring that third-parties are not accessing data to which they are not authorized. A query may also undergo anonymisation techniques including the addition of noise or substitution to further protect privacy. It is still an open research issue to find a good interactive query-response mechanism which is at the same time capable of answering any questions fairly accurately (meaning in the less noisy way), while simultaneously preserving privacy.

To limit inference and linkability attacks it is necessary to keep track of the queries issued by an entity and to observe the information gained about data subjects; accordingly, “differential-privacy” databases should not be deployed on open search engines that offer no traceability of the querying entities.

3.1.3.1 Guarantees

- *Singling out:* If only statistics are output and the rules applied to the set are well chosen, it should not be possible to use the answers to single out an individual.
- *Linkability:* By using multiple requests it might be possible to link the entries relating to a specific individual between two answers.

---

[14] Dwork, C. (2006). Differential privacy. In *Automata, languages and programming* (pp. 1-12). Springer Berlin Heidelberg.

[15] Cf. Ed Felten (2012) Protecting privacy by adding noise. URL: https://techatftc.wordpress.com/2012/06/21/protecting-privacy-by-adding-noise/.

- *Inference:* It is possible to infer information about individuals or groups by using multiple requests.

3.1.3.2. Common mistakes

- Not injecting enough noise: To prevent linking with background knowledge, the challenge is to provide minimal evidence about whether or not a specific data subject or a group of data subjects contributed to the data set. The major difficulty from a data protection perspective is to be able to generate the proper amount of noise to add to the true answers, so as to protect individuals' privacy while preserving the usefulness of the released responses.

3.1.3.3 Failures of differential privacy

Treating each query independently: A combination of query results may allow disclosing information which was intended to be secret. If a query history is not retained, then an attacker may engineer multiple questions to a "differential privacy" database that progressively reduce the amplitude of the outputted sample until a specific character of a single data subject or a group of data subjects might emerge, deterministically or with very high likelihood. Furthermore, an additional caveat is to avoid the mistake of thinking the data are anonymous for the third party, while the data controller can still identify the data subject in the original database taking into account all the means likely reasonably to be used.

3.2. Generalization

Generalization is the second family of anonymisation techniques. This approach consists of generalizing, or diluting, the attributes of data subjects by modifying the respective scale or order of magnitude (i.e. a region rather than a city, a month rather than a week). Whilst generalization can be effective to prevent singling out, it does not allow effective anonymisation in all cases; in particular, it requires specific and sophisticated quantitative approaches to prevent linkability and inference.

3.2.1. Aggregation and K-anonymity

Aggregation and K-anonymity techniques aim to prevent a data subject from being singled out by grouping them with, at least, k other individuals. To achieve this, the attribute values are generalized to an extent such that each individual shares the same value. For example, by lowering the granularity of a location from a city to a country a higher number of data subjects are included. Individual dates of birth can be generalized into a range of dates, or grouped by month or year. Other numerical attributes (e.g. salaries, weight, height, or the dose of a medicine) can be generalized by interval values (e.g. salary €20,000 – €30,000). These methods may be used when the correlation of punctual values of attributes may create quasi-identifiers.

3.2.1.1. Guarantees

- *Singling out:* Because the same attributes are now shared by k users, it should be no longer possible to single out an individual within a group of k users.
- *Linkability:* While linkability is limited, it remains possible to link records by groups of k users. Then, within this group, the probability that two records correspond to the

same pseudo-identifiers is 1/k (which might be significantly higher than the probability that such entries are unlinkable).

- *Inference:* The main flaw of the k-anonymity model is that it does not prevent any type of inference attack. Indeed, if all k individuals are within a same group, then if it is known which group an individual belongs to, it is trivial to retrieve the value of this property.

3.2.1.2. Common mistakes

- Missing some quasi-identifiers: A critical parameter when considering k-anonymity is the threshold of k. The higher the value of k, the stronger the privacy guarantees. A common mistake is to artificially augment the value k by reducing the considered set of quasi-identifiers. Reducing quasi-identifiers makes it easier to build clusters of k-users due to the inherent power of identification associated to the other attributes (especially if some of them are sensitive or possess a very high entropy, as in the case of very rare attributes). Not considering all the quasi-identifiers when selecting the attribute to generalize is a critical mistake; if some attributes can be used to single out an individual in a cluster of k, then the generalization fails to protect some individuals (see example in Table 2).
- Small value of k: Aiming for a small value of k is similarly problematic. If k is too small, the weight of any individual in a cluster is too significant and inference attacks have a higher success rate. For instance, if k=2 then the probability for the two individuals to share the same property is higher than for k>10.
- Not grouping individuals with the same weight: Grouping a set of individuals with an uneven distribution of attributes can also be problematic. The impact of an individual's record on a dataset will vary: some will represent a significant fraction for the entries while the contributions of others remain fairly insignificant. It is therefore important to make sure that k is high enough so that no individuals represent too important a fraction of the entries in a cluster.

3.1.3.3. Failures of k-anonymity

The main issue with k-anonymity is that it does not prevent inference attacks. In the following example, if the attacker knows that a specific individual is in the data set and was born in 1964, he also knows that the individual had a heart attack Furthermore, if we know that this dataset was obtained from a French organisation then each individual resides in Paris as the first three digits of Parisian postcodes are 750*).

Year	Gender	ZIP	Diagnosis
1957	M	750*	Heart attack
1957	M	750*	Cholesterol
1957	M	750*	Cholesterol
1964	M	750*	Heart attack
1964	M	750*	Heart attack

Table 2. An example of poorly engineered k-anonymisation

3.2.2. L-diversity/T-closeness

L-diversity extends k-anonymity to ensure that deterministic inference attacks are no longer possible by making sure that in each equivalence class every attribute has at least *l* different values.

One basic goal to achieve is to limit the occurrence of equivalence classes with poor attribute variability, so that an attacker with background knowledge on a specific data subject is always left with a significant uncertainty.

L-diversity is useful to protect data against inference attacks when the values of attributes are well distributed. It has to be highlighted, however, that this technique cannot prevent the leakage of information if the attributes within a partition are unevenly distributed or belong to a small range of values or semantic meanings. In the end, l-diversity is subject to probabilistic inference attacks.

T-closeness is a refinement of l-diversity, in that it aims to create equivalent classes that resemble the initial distribution of attributes in the table. This technique is useful when it is important to keep the data as close as possible to the original one; to that end, a further constraint is placed on the equivalence class, namely that not only at least *l* different values should exist within each equivalence class, but also that each value is represented as many times as necessary to mirror the initial distribution of each attribute.

3.2.2.1. Guarantees

- *Singling out:* Like k-anonymity, l-diversity and t-closeness can ensure that records relating to an individual cannot be singled out in the database.
- *Linkability:* l-diversity and t-closeness are not an improvement over k-anonymity with regard to unlinkability. The issue is the same as with any cluster: the probability that the same entries belong to a same data subject is higher than 1/N (where N is the number of data subjects in the database).
- *Inference:* The main improvement of l-diversity and t-closeness over k-anonymity is that it is no longer possible to set up inference attacks against a ‘l-diverse’ or “t close” database with a 100% confidence.

3.2.2.2. Common mistakes

- Protecting sensitive attribute values by mixing them with other sensitive attributes: It is not enough to have two values of an attribute in a cluster in order to provide privacy guarantees. In fact, the distribution of sensitive values in each cluster should resemble the distribution of those values in the total population, or at least it should be uniform throughout the cluster.

3.2.2.3. Failures of l-diversity

In the table below, l-diversity is granted with respect to the attribute “Diagnosis”; however, knowing that an individual born in 1964 is in this table, it is still possible to assume with a very high probability that he had a heart attack.

Year	Gender	ZIP	Diagnosis
1957	M	750*	**Heart attack**
1957	M	750*	**Cholesterol**
1957	M	750*	**Cholesterol**
1957	M	750*	**Cholesterol**
1964	M	750*	**Heart attack**
1964	M	750*	**Heart attack**
1964	M	750*	**Heart attack**
1964	M	750*	**Cholesterol**
1964	M	750*	**Heart attack**
1964	M	750*	**Heart attack**
1964	M	750*	**Heart attack**
1964	M	750*	**Heart attack**
1964	M	750*	**Heart attack**
1964	M	750*	**Heart attack**
1964	M	750*	**Heart attack**

Table 3. An l-diverse table where “Diagnosis” values are not uniformly distributed

Name	Date of birth	Gender
Smith	1964	M
Rossi	1964	M
Dupont	1964	M
Jansen	1964	M
Garcia	1964	M

Table 4. Knowing that these individuals are in Table 3, an attacker could infer that they had a heart attack

4. Pseudonymisation

Pseudonymisation consists of replacing one attribute (typically a unique attribute) in a record by another. The natural person is therefore still likely to be identified indirectly; accordingly, pseudonymisation when used alone will not result in an anonymous dataset. Nevertheless, it is discussed in this opinion because of the many misconceptions and mistakes surrounding its use.

Pseudonymisation reduces the linkability of a dataset with the original identity of a data subject; as such, it is a useful security measure but not a method of anonymisation.

The result of pseudonymisation can be independent of the initial value (as is the case of a random number generated by the controller or a surname chosen by the data subject) or it can be derived from the original values of an attribute or set of attributes e.g. a hash function or encryption scheme.

The most used pseudonymisation techniques are as follows:

- Encryption with secret key: in this case, the holder of the key can trivially re-identify each data subject through decryption of the dataset because the personal data are still contained in the dataset, albeit in an encrypted form. Assuming that a state-of-the-art encryption scheme was applied, decryption can only be possible with the knowledge of the key.
- Hash function: this corresponds to a function which returns a fixed size output from an input of any size (the input may be a single attribute or a set of attributes) and cannot be reversed; this means that the reversal risk seen with encryption no longer exists. However, if the range of input values the hash function are known they can be replayed through the hash function in order to derive the correct value for a particular record. For instance, if a dataset was pseudonymised by hashing the national identification number, then this can be derived simply by hashing all possible input values and comparing the result with those values in the dataset. Hash functions are usually designed to be relatively fast to compute, and are subject to brute force attacks.[16] Pre-computed tables can also be created to allow for the bulk reversal of a large set of hash values.

  The use of a salted-hash function (where a random value, known as the "salt", is added to the attribute being hashed) can reduce the likelihood of deriving the input value but nevertheless, calculating the original attribute value hidden behind the result of a salted hash function may still be feasible with reasonable means.[17]
- Keyed-hash function with stored key: this corresponds to a particular hash function which uses a secret key as an additional input (this differs from a salted hash function as the salt is commonly not secret). A data controller can replay the function on the attribute using the secret key, but it is much more difficult for an attacker to replay the function without knowing the key as the number of possibilities to be tested is sufficiently large as to be impractical.

---

[16] Such attacks consist in trying all the plausible inputs in order to build correspondence tables.

[17] Especially if the type of attribute is known (name, social number, date of birth, etc). To add computational requirement, one could rely on a key derivation hash function, where the computed value is hashed several times with a short salt.

- Deterministic encryption or keyed-hash function with deletion of the key: this technique may be equated to selecting a random number as a pseudonym for each attribute in the database and then deleting the correspondence table. This solution allows[18] diminishing the risk of linkability between the personal data in the dataset and those relating to the same individual in another dataset where a different pseudonym is used. Considering a state-of-the-art algorithm, it will be computationally hard for an attacker to decrypt or replay the function, as it would imply testing every possible key, given that the key is not available.

- Tokenization: this technique is typically applied in (even if it is not limited to) the financial sector to replace card ID numbers by values that have reduced usefulness for an attacker. It is derived from the previous ones being typically based on the application of one-way encryption mechanisms or the assignment, through an index function, of a sequence number or a randomly generated number that is not mathematically derived from the original data.

4.1. Guarantees

- *Singling out:* It is still possible to single out individuals' records as the individual is still identified by a unique attribute which is the result of the pseudonymisation function (= the pseudonymised attribute).

- *Linkability:* Linkability will still be trivial between records using the same pseudonymised attribute to refer to the same individual. Even if different pseudonymised attributes are used for the same data subject, linkability may still be possible by means of other attributes. Only if no other attribute in the dataset can be used to identify the data subject and if every link between the original attribute and the pseudonymised attribute has been eliminated (including by deletion of the original data), will there be no obvious cross-reference between two datasets using different pseudonymised attributes.

- *Inference:* Inference attacks on the real identity of a data subject are possible within the dataset or across different databases that use the same pseudonymised attribute for an individual, or if pseudonyms are self-explanatory and do not mask the original identity of the data subject properly.

4.2. Common mistakes

- Believing that a pseudonymised dataset is anonymised: Data controllers often assume that removing or replacing one or more attributes is enough to make the dataset anonymous. Many examples have shown that this is not the case; simply altering the ID does not prevent someone from identifying a data subject if quasi-identifiers remain in the dataset, or if the values of other attributes are still capable of identifying an individual. In many cases it can be as easy to identify an individual in a pseudonymised dataset as with the original data. Extra steps should be taken in order to consider the dataset as anonymised, including removing and generalising attributes or deleting the original data or at least bringing them to a highly aggregated level.

- Common mistakes when using pseudonymisation as a technique to reduce linkability:

[18] Depending on the other attributes in the dataset and on the deletion of the original data.

- o <u>Using the same key in different databases:</u> eliminating the linkability of different datasets depends highly on the use of a keyed algorithm and the fact that a single individual will correspond to different pseudonymised attributes in different contexts. It is thus important to avoid using the same key in different databases to be able to reduce linkability.
- o <u>Using different keys ("rotating keys") for different users:</u> it might be tempting to use different keys for different sets of users and to change the key on a per use basis (for instance, use the same key to record 10 entries relating to the same user). However, if not properly engineered this operation might trigger the occurrence of patterns, partially reducing the intended benefits. For instance, rotating the key by way of specific rules for specific individuals would facilitate linkability of the entries corresponding to a given individuals. Also, the disappearance of a recurrent pseudonymised data in the database at the time a new one appears may signal that both records relate to the same natural person.
- o <u>Keeping the key:</u> if the secret key is stored alongside the pseudonymised data, and the data are compromised, then the attacker may be able to trivially link the pseudonymised data to their original attribute. The same applies if the key is stored separately from the data but not in a secure manner.

4.3. Shortcomings of Pseudonymisation

- *<u>Health care</u>*

1. Name, address date of birth	**2. Period of Special Assistance Benefit.**	**3. Body mass index**	**6. Research cohort reference no.**
[redacted]	< 2 years	15	QA5FRD4
	> 5 years	14	2B48HFG
	< 2 years	16	RC3URPQ
	> 5 years	18	SD289K9
	< 2 years	20	5E1FL7Q

Table 5. An example of pseudonymisation by hashing (name, address date of birth) which can be easily reversed

A dataset has been created to examine the relationship between a person's weight and the receipt of a special assistance benefit payment. The original dataset included the data subjects' name, address and date of birth but this has been deleted. The research cohort reference number was generated from the deleted data using a hash function. Although the name, address and date of birth were deleted from the table, if a data subject's name, address and date of birth is known in addition to knowing the hash function used it is easy to calculate the research cohort reference numbers.

- *<u>Social Networks</u>*

It has been shown[19] that sensitive information about specific individuals can be extracted from social-network graphs, in spite of the "pseudonymisation" techniques applied to such data. A provider of a social network assumed wrongly that pseudonymisation was

[19] A. Narayanan and V. Shmatikov, "De-anonymizing social networks," in 30th IEEE Symposium on Security and Privacy, 2009.

robust to prevent identification after selling the data to other companies for marketing and advertising purposes. Instead of real names, the provider used nicknames, but this clearly was not enough to anonymise user profiles, since the relationships between the different individuals are unique and can be used as an identifier.

- *Locations*

Researchers at MIT[20] recently analyzed a pseudonymised dataset consisting of 15 months of spatial-temporal mobility coordinates of 1,5 million people on a territory within a radius of 100 km. They showed that 95% of the population could be singled-out with four location points, and that just two points were enough to single-out more than 50% of the data subjects (one of such points is known, being very likely “home” or “office”) with very limited space for privacy protection, even if the individuals’ identities were pseudonymised by replacing their true attributes [….] with other labels.

5. Conclusions & Recommendations

5.1. Conclusions

Techniques of de-identification and anonymisation are the subject of intense research, and this paper has shown consistently that each technique has its advantages and disadvantages. In most cases it is not possible to give minimum recommendations for parameters to use as each dataset needs to be considered on a case-by-case basis.

In many cases, an anonymised dataset can still present residual risk to data subjects. Indeed, even when it is no longer possible to precisely retrieve the record of an individual, it may remain possible to glean information about that individual with the help of other sources of information that are available (publicly or not). It has to be highlighted that beyond the direct impact on data subjects produced by the consequences of a poor anonymisation process (annoyance, time consumption and feeling of lost control by being included in a cluster without awareness or prior consent), other indirect side effects of poor anonymisation may occur whenever a data subject is included in a target erroneously by some attacker, as a consequence of processing anonymised data - especially if the attacker’s intents are malicious. Therefore the Working Party stresses that anonymisation techniques can provide privacy guarantees, but only if their application is engineered appropriately – which means that the prerequisites (context) and the objective(s) of the anonymisation process must be clearly set out in order to achieve the targeted anonymisation level.

5.2. Recommendations

- Some anonymisation techniques show inherent limitations. These limitations must be considered seriously before a given technique is used to craft an anonymisation process by data controllers. They must have regard to the purposes to be achieved through anonymisation – such as protecting individuals’ privacy when publishing a dataset, or allowing a piece of information to be retrieved from a dataset.
- Each technique described in this paper fails to meet with certainty the criteria of effective anonymisation (i.e. no singling out of an individual; no linkability between records relating to an individual; and no inference concerning an individual). However as some of these risks may be met in whole or in part by a given technique, careful engineering is

[20] Y.-A. de Montjoye, C. Hidalgo, M. Verleysen and V. Blondel, "Unique in the Crowd: The privacy bounds of human mobility," Nature, no. 1376, 2013.

necessary in devising the application of an individual technique to the specific situation and in applying a combination of those techniques as a way to enhance the robustness of the outcome.

The table below provides an overview of the strengths and weakness of the techniques considered in terms of the three basic requirements:

	Is Singling out still a risk?	**Is Linkability still a risk?**	**Is Inference still a risk?**
Pseudonymisation	Yes	Yes	Yes
Noise addition	Yes	May not	May not
Substitution	Yes	Yes	May not
Aggregation or K-anonymity	No	Yes	Yes
L-diversity	No	Yes	May not
Differential privacy	May not	May not	May not
Hashing/Tokenization	Yes	Yes	May not

Table 6. Strengths and Weaknesses of the Techniques Considered

- The optimal solution should be decided on a case-by-case basis. A solution (i.e. a complete anonymisation process) meeting the three criteria would be robust against identification performed by the most likely and reasonable means the data controller or any third party may employ.
- Whenever a proposal does not meet one of the criteria, a thorough evaluation of the identification risks should be performed. This evaluation should be provided to the authority if national law requires that the authority shall assess or authorise the anonymisation process.

To reduce the identification risks, the following good practices should be taken into consideration:

<u>Good anonymisation practices</u>

In general:

- Do not rely on the "release and forget" approach. Given the residual risk of identification, data controllers should:
  - 1. Identify new risks and re-evaluate the residual risk(s) regularly,
  - 2. Assess whether the controls for identified risks suffice and adjust accordingly; AND
  - 3. Monitor and control the risks.
- As part of such residual risks, take into account the identification potential of the non-anonymised portion of a dataset (if any), especially when combined with the anonymised portion, plus of possible correlations between attributes (e.g. between geographical location and wealth level data).

Contextual elements:

- The purposes to be achieved by way of the anonymised dataset should be clearly set out as they play a key role in determining the identification risk.
- This goes hand in hand with the consideration of all the relevant contextual elements – e.g., nature of the original data, control mechanisms in place (including security measures to restrict access to the datasets), sample size (quantitative features), availability of public information resources (to be relied upon by the recipients), envisaged release of data to third parties (limited, unlimited e.g. on the Internet, etc.).
- Consideration should be given to possible attackers by taking account of the appeal of the data for targeted attacks (again, sensitivity of the information and nature of the data will be key factors in this regard).

Technical elements:

- Data controllers should disclose the anonymisation technique / the mix of techniques being implemented, especially if they plan to release the anonymised dataset.
- Obvious (e.g. rare) attributes / quasi-identifiers should be removed from the dataset.
- If noise addition techniques are used (in randomization), the noise level added to the records should be determined as a function of the value of an attribute (that is, no out-of-scale noise should be injected), the impact for data subjects of the attributes to be protected, and/or the sparseness of the dataset.
- When relying on differential privacy (in randomization), account should be taken of the need to keep track of queries so as to detect privacy-intrusive queries as the intrusiveness of queries is cumulative.
- If generalization techniques are implemented, it is fundamental for the data controller not to limit themselves to one generalization criterion even for the same attribute; that is to say, different location granularities or different time intervals should be selected. The selection of the criterion to be applied must be driven by the distribution of the attribute values in the given population. Not all distributions lend themselves to being generalized – i.e., no one-size-fits-all approach can be followed in generalization. Variability within equivalence classes should be ensured; for instance, a specific threshold should be selected depending on the "contextual elements" mentioned above (sample size, etc.) and if that threshold is not reached, then the specific sample should be discarded (or a different generalization criterion should be set).

ANNEX

A primer on anonymisation techniques

A.1. Introduction

Anonymity is interpreted differently across the EU – corresponding in some countries to computational anonymity (i.e., it should be computationally difficult, even for the controller in collaboration with any party, to identify directly or indirectly one of the data subjects) and in other countries to perfect anonymity (i.e. it should be impossible, even for the controller in collaboration with any party, to identify directly or indirectly one of the data subjects). Nevertheless, "anonymisation" corresponds in both cases to the process by which data are made anonymous. The difference lies in what is considered as an acceptable level for the risk of re-identification.

Various use cases can be envisaged for anonymised data, ranging from social surveys, statistical analyses, new service/product development. Sometimes, even such general purpose activities may have an impact on specific data subjects, nullifying the supposedly anonymous nature of the processed data. Many examples can be given, from the launch of targeted marketing initiatives, to the implementation of public measures based on user profiling, or behaviours, or mobility patterns[21].

Unfortunately, beyond general statements no mature metric exists to evaluate in advance the time or effort necessary for re-identification after the processing, or alternatively to select the most appropriate procedure to put in place if one wishes to lower the probability that a released database refers to an identified set of data subjects.

The "art of anonymisation", as these practices are sometimes referred to in scientific literature[22], is a new scientific branch which is still in its infancy and many practices exist to degrade the power of identification of data sets; however, it has to be clearly stated that the majority of such practices do not prevent linking the processed data with data subjects. In some circumstances identification of datasets considered anonymous has been proven to be very successful, in other situations false positives have occurred.

Broadly speaking two different approaches exist: one is based on attribute generalization, another one on randomization. Going through the details and subtleties of these practices will bring us to a new insight on the power of identification of data, and will shed a new light on the very notion of personal data.

A.2. "Anonymisation" by randomization

One option for anonymisation consists of modifying the actual values to prevent linking between the anonymised data and the original values. This goal can be reached through a wide number of methodologies ranging from noise injection to data swapping (permutation). It has to be stressed that removing an attribute is equivalent to an extreme form of randomization of this attribute (being the attribute entirely covered by noise).

In some circumstances, the goal of the overall processing is not so much in releasing a randomized data set, but rather to grant access to the data by means of queries. The risk for the data subject in this case stems from the likelihood of an attacker being able to withdraw information from a series of different queries, without the knowledge of the data controller. In

[21] For example the case of TomTom in the Netherlands (see example explained in paragraph 2.2.3).

[22] Jun Gu,Yuexian Chen,Junning Fu,HuanchunPeng,Xiaojun Ye, Synthesizing: Art of Anonymization, Database and Expert Systems Applications Lecture Notes in Computer Science –Springer- Volume 6261, 2010, pp 385-399

order to guarantee anonymity to individuals in the dataset about it should not be possible to conclude that a data subject contributed to the data set, thus breaking the link with any kind of background information an attacker may have.

Adding noise as appropriate to the query response, can further reduce the risk of re-identification. This approach, also known in literature as differential privacy[23], departs from those described earlier in that it gives data publishers greater control over access to the data as compared to public release. The addition of noise has two main objectives: one, to protect the privacy of data subjects in the dataset and, two, retain the usefulness of the released information. In particular the magnitude of noise must be proportionate with the level of querying (too many queries on individuals to be answered too accurately result increasing the likelihood of identification). Today the successful application of randomization must be considered on a case-by-case basis with no technique offering a fool proof methodology as examples exist of information leaks on the attributes of a data subject (being this included or not in the data set) even when the dataset was considered randomized by the data controller.

It may be helpful to discuss specific examples to clarify the potential failures of randomization as a means to provide anonymisation. For instance, in the context of interactive access, queries considered as privacy friendly might pose a risk to the data subjects. In fact, if the attacker knows that a subgroup S of individuals is within the dataset which contains information on the incidence of attribute A within a population P, by simply querying with the two questions "How many individuals in population P possess attribute A?" and "How many individuals in population P, except those belonging to subgroup S, possess attribute A?" it may be possible to determine (by difference) the number of individuals in S who actually possess attribute A - either deterministically or by likelihood inference. In any case, the privacy of the individuals in subgroup S might be seriously jeopardised, especially depending on the nature of attribute A.

It can also be considered that if a data subject is not in the dataset but his relationship to data within the dataset is known then release of the dataset may cause a risk to his privacy. For example, if it is known that "the target's value of attribute A differs by a quantity X from the average value of the population", by simply asking the database curator to perform the privacy-friendly operation of extracting the average value of attribute A, the attacker may exactly infer a personal data relating to a specific data subject.

Injecting some relative inaccuracies into the actual values in a database is an operation that should be properly engineered. Sufficient noise needs to be added to protect privacy but also small enough to preserve data usefulness. For instance, if the number of data subjects with a peculiar attribute is very small or the sensitiveness of the attribute high it can be better to report a range or a generic sentence such as "a small number of cases, possibly even zero", instead of reporting the actual number. In this way, even if the noisy disclosure mechanism is known in advance, the privacy of the data subject is preserved, since a degree of uncertainty remains. From an usefulness perspective, if the inaccuracy is properly engineered, the results are still useful for statistical or decision-making purposes.

Database randomization, and differential privacy access require further thought. Firstly, the right amount of distortion may vary significantly according to the context (the type of query, the size of the population in the database, the nature of the attribute and its inherent power of identification) and no "*ad omnia*" solution can be envisaged. Furthermore, the context may

[23] Cynthia Dwork, Differential Privacy, International Colloquium on Automata, Languages and Programming (ICALP) 2006, p. 1–12

change over time and the interactive mechanism should be modified accordingly. Calibrating the noise requires tracking the cumulative privacy risks which any interactive mechanism poses to data subjects. The data access mechanism should then be equipped with alerts when a “privacy cost” budget has been reached and data subjects might be exposed to specific risks if a new query is set forth, to assist the data controller in determining the appropriate level of distortion to inject every time into the actual personal data.

On the other hand, we should also consider the case where the values of attributes are deleted (or modified). A commonly used solution to deal with some atypical values for attributes is the deletion either of the set of data related to the atypical individuals or of the atypical values. In the latter case, it is then important to assure than the absence of value in itself does not become an element for the identification of a data subject.

Let us now consider randomization by attribute substitution. A major misconception when dealing with anonymisation is equating it to encryption or key-coding. This misconception is grounded in two assumptions – namely, a) that once encryption is applied to some attributes of a record in a database (e.g. name, address, birth date), or these attributes are substituted by a seemingly randomized string as a result of a key-coding operation like keyed hash-function, then that record is “anonymised”, and b) that anonymisation is more effective if the length of the key is appropriate and the encryption algorithm is state of the art. This misconception is widespread among data controllers and deserves clarification, as also related to pseudonymisation and its allegedly lower risks.

First of all, the goals of these techniques are radically different: encryption as a security practice aims to provide the confidentiality of a communication channel between identified parties (human beings, devices, or pieces of software/hardware) to avoid eavesdropping or unintended disclosure. Key-coding corresponds to a semantic translation of the data depending on a secret key. On the other hand, the goal of anonymisation is avoiding identification of individuals by preventing hidden linking of attributes to a data subject.

Neither encryption nor key-coding per se lends itself to the goal of making a data subject unidentifiable: as, in the hands of the controller at least, the original data are still available or deducible. The sole implementation of a semantic translation of personal data, as happens with key-coding, does not eliminate the possibility to restore the data back to their original structure - either by applying the algorithm in the opposite way, or by brute force attacks, depending on the nature of the schemes, or as a result of a data breach. State-of-the-art encryption can ensure that data is protected to a higher degree, i.e. it is unintelligible for entities that ignore the decryption key, but it does not necessarily result in anonymisation. For as long as the key or the original data are available (even in the case of a trusted third party, contractually bound to provide secure key escrow service), the possibility to identify a data subject is not eliminated. .

Focusing only on the robustness of the encryption mechanism as a measure of the degree of “anonymisation” of a dataset is misleading as many other technical and organizational factors affect the overall security of an encryption mechanism or hash function. Many successful attacks have been reported in the literature which entirely bypass the algorithm, either because they leverage on weakness in the custody of the keys (e.g. the existence of a less secure default mode) or other human factors (e.g. weak passwords for key recovery). Finally, a chosen encryption scheme with a given key size is designed to ensure confidentiality for a given period (most of the current keys will have to be resized around 2020), while an anonymisation process should not be limited in time.

It is worth elaborating now on the limits of attribute randomization (or substitution and removal) taking into account various bad examples of anonymisation by randomization occurred in the recent years, and the reasons behind such failures.

A well-known case involving the release of a poorly anonymised data set is that of the Netflix prize[24]. Looking at a generic record in a database where a number of attributes relating to a data subject have been randomized, each record can be still be partitioned into two sub-records as follows: {randomized attributes, clear attributes}, where the clear attributes may be any combination of supposedly non personal data. A specific observation which can be made from the Netflix Prize dataset comes from the consideration that each record can be represented by a point in a multidimensional space, where each clear attribute is a coordinate. Using this technique any dataset can be regarded as a constellation of points in such a multi-dimensional space which can exhibit a high degree of sparseness, meaning that points are can be distant from one another. Indeed, they can be so distant that after partitioning the space into wide regions, each region contains only one record. Even the injection of noise fails to being records sufficiently close together to share that same multi-dimensional region. For instance, in the Netflix experiment, records were sufficiently unique with just 8 movie ratings given within 14 days of distance. After the addition of noise to both the ratings and dates no superposition of regions could be observed. In other words, the very same selection of just 8 rated movies constituted a fingerprint of the expressed ratings, not shared between two data subjects within the database. Based on this geometric observation, the researchers matched the supposedly anonymous Netflix data set with another public database with movie ratings (the IMDB), thus finding users who had expressed ratings for the same movies within the same time intervals. Since the majority of users showed a one-to-one correspondence the auxiliary information retrieved in the IMDB database could be imported into the released Netflix data set, thus enriching with identities all the supposedly anonymised records.

It is important to stress that this is a general property: the residual part of any “randomized” database still possesses a very high power of identification, depending on the rarity of the combination of the residual attributes. This is a caveat that data controllers should always bear in mind when selecting randomization as their way to reach targeted anonymisation.

Many re-identification experiments of this type have followed a similar approach of projecting two databases onto the same subspace). This is a very powerful re-identification methodology, which had recently many applications in different areas. For example an identification experiment performed against a social network[25], exploited the social graph of users pseudonymised by means of labels. In this case, the attributes used for identification were the list of contacts of each user as it was shown that the likelihood of an identical list of contacts between two individuals is very low. Based on this intuitive assumption, it has been found that a sub-graph of the internal connections of a very limited number of nodes constitutes a topological fingerprint to retrieve, hidden within the network, and that a wide portion of the entire social network can be identified once this sub-network has been identified. Just to provide some figures on the performances of a similar attack, it was shown that by using fewer than 10 nodes (which can give rise to million different sub-network configurations, each potentially constituting a topological fingerprint) a social network of more than 4 million pseudonymised nodes and 70 million links can be prone to re-

[24] Arvind Narayanan, Vitaly Shmatikov: Robust De-anonymization of Large Sparse Datasets. IEEE Symposium on Security and Privacy 2008: 111-125

[25] L. Backstrom, C. Dwork, and J. M. Kleinberg. Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography, Proceedings of the 16th International Conference on World Wide Web WWW'07, page 181-190. (2007)

identification attacks, and the privacy of a high number of connections can be compromised. It has to be highlighted that this re-identification approach is not tailored to the specific context of social networks, but it is general enough to be potentially adapted to other databases where relationships between users are recorded (e.g. phone contacts, email correspondence, dating sites etc.).

Another way to identify a supposedly anonymous record is based on the analysis of writing style (stylometry)[26]. A number of algorithms have already been developed to extract metrics from parsed text including the frequency of particular word usage, the occurrence of specific grammatical patterns and the type of punctuation. All of these properties can be used to anchor a supposedly anonymous text to the writing style of an identified author. Researchers have retrieved the writing style of more than 100.000 blogs, and are today capable of automatically identifying the author of a post with a precision score already approaching 80%; the accuracy of this technique is expected to grow further also exploiting other signals, such as location or other metadata contained in the text.

The power of identification using the semantics of a record (i.e. the residual non randomized part of a record) is an issue deserving more consideration by the research community and by industry. The recent reversal of the identities of DNA donors (2013)[27] shows that very little progress has been made since the well-known AOL incident (2006) – when a database containing twenty million search keywords for over 650,000 users over a 3-month period was publicly released. This resulted into the identification and location of a number of AOL users.

Another family of data which is rarely anonymised just by removing the identities of data subjects or by partially encrypting some attributes are location data. The mobility patterns of human beings are may be sufficiently unique that the semantic part of location data (the places where the data subject was at a specific time), even without other attributes, is capable of revealing many traits of a data subject[28]. This was proven many times in representative academic studies[29].

In this regard it is necessary to warn on the use of pseudonyms as a way to afford adequate protection to data subjects against identity or attribute leaks. If pseudonymisation is based on the substitution of an identity by another unique code, the presumption that this constitutes a robust de-identification is naïf and does not take into account the complexity of identification methodologies and the multifarious contexts where they might be applied.

A.3. "Anonymisation" by generalization

A simple example may help clarify the approach based on attribute generalization.

---

[26] http://33bits.org/2012/02/20/is-writing-style-sufficient-to-deanonymize-material-posted-online/

[27] Genetic data are an especially significant example of sensitive data that can be at risk of re-identification if the sole mechanism to "anonymize" them is the removal of donors' identities. See the example quoted in paragraph 2.2.2 above. See also John Bohannon, Genealogy Databases Enable Naming of Anonymous DNA Donors, Science, Vol. 339, No. 6117 (18 January 2013), pp. 262

[28] This issue has been tackled in some national legislations. For instance, in France, published location statistics are anonymized through generalization and permutation. Hence, INSEE publish statistics that are generalized by aggregating all data to a 40000 square meters area. The granularity of the dataset is enough to preserve data utility and permutations prevent deanonymisation attacks in sparse areas. More generally, aggregating this family of data and permuting them provides strong guarantees against inference and de-anonymisation attacks (http://www.insee.fr/en/).

[29] de Montjoye, Y.-A., Hidalgo, C.A., Verleysen, M. & Blondel, V.D. Unique in the Crowd: The privacy bounds of human mobility. Nature. 3, 1376 (2013)

Let us consider the case where a data controller decides to release a simple table containing three pieces of information, or attributes: an identification number, unique to each record, a location identification, which links the data subject to the place where he lives, and a property identification, which shows the property that data subject has; let us further assume that this property is one of two distinct values, generically indicated by {P1, P2} :

Serial ID	Location ID	Property
#1	Rome	P1
#2	Madrid	P1
#3	London	P2
#4	Paris	P1
#5	Barcelona	P1
#6	Milan	P2
#7	New York	P2
#8	Berlin	P1

Table A1. Sample of data subjects gathered by location and properties P1 and P2

If someone, referred to as the attacker, knows in advance that a specific data subject (the target) who lives in Milan is included in the table, then after inspecting the table he can learn that, being #6 the only data subject with that Location ID, he also possesses property P2.

This very basic example shows the main elements of any identification procedure applied to a set of data that has gone through a supposedly anonymisation process. Namely, there is an attacker who (accidentally or intentionally) possessed background knowledge on some or all data subjects in a dataset. The attacker aims at linking this background knowledge with the data in the released dataset to get a clearer picture of the characteristics of those data subjects.

In order to make data linking with any kind of background knowledge less effective or less immediate, the data controller could focus on the Location ID, substituting the city where the data subjects live by a wider area like the Country. In this way the table would look as follows.

Serial ID	Location ID	Property
#1	Italy	P1
#2	Spain	P1
#3	UK	P2
#4	France	P1
#5	Spain	P1
#6	Italy	P2
#7	USA	P2
#8	Germany	P1

Table A2. Generalization of table A1 by nationality

With this new data aggregation, the attacker's background knowledge on an identified data subject (say, "the target lives in Rome and he is in the table") does not allow drawing any clear conclusion on his property: this is because the two Italians in the table have distinct properties, P1 and P2 respectively. The attacker is left with a 50% uncertainty on the property of the target entity. This simple example shows the effect of generalization on the practice of anonymisation. In fact, while this generalization trick might be effective to halve the

probability to identify an Italian target, it is not effective for a target from other locations (e.g. USA).

Furthermore, an attacker may still learn information on a Spanish target. If the background knowledge is of the type “the target lives in Madrid and he is in the table” or “the target lives in Barcelona and he is in the table”, the attacker can infer with 100% certainty that the target possesses property P1. Therefore generalization does not yield the same level of privacy or resistance against inference attacks to the whole population in the dataset.

Following this line of reasoning, one might be tempted to conclude that stronger generalization might be helpful to prevent any linking - for instance a generalization by Continent. In this way the table would look as follows

Serial ID	Location ID	Property
#1	Europe	P1
#2	Europe	P1
#3	Europe	P2
#4	Europe	P1
#5	Europe	P1
#6	Europe	P2
#7	North America	P2
#8	Europe	P1

Table A3. Generalization of table A1 by continent

With this kind of aggregation, all data subjects in the table, except the one living in the US, would be protected against linking and identification attacks, and any background information of the type “the target lives in Madrid and he is in the table” or “the target lives in Milan and he is in the table” would lead to some level of likelihood as for the property applying to the given data subject (P1 with probability 71,4% and P2 with probability 28,6%), rather than to direct linking. Also, this further generalization comes at the expense of an evident and radical loss of information: the table does not allow discovering potential correlations between the properties and the location, namely whether a specific location might trigger any of the two properties with higher probability, as it only yields the so called “marginal” distributions, namely the absolute probability of occurrence of property P1 and P2 population-wide (respectively 62,5% and 37,5% in our example) and within each Continent (respectively, as said,71,4% and 28,6% in Europe and 100% and 0% in North America).

The example also shows that the practice of generalization affects the practical usefulness of data. Some engineering tools are available as of today to tackle in advance (i.e., before a dataset is released) what the most appropriate level of attribute generalization is, so as to reduce the risks of identification for the data subjects within a table without affecting usefulness of the released data to an excessive degree.

k-anonymity

An attempt to prevent linking attacks, based on the generalization of attributes, is known as k-anonymity. This practice stems from a re-identification experiment performed in the late 1990’s, where a private U.S. company, active in the health sector, publicly released a supposedly anonymised dataset. This anonymisation consisted of removing data subjects’ names, but the dataset still contained health data and other attributes such as the ZIP code (the

location ID where they lived), gender, and the full date of birth. The same triplet {ZIP code, gender, full date of birth} was also included also in other publicly available registers (e.g., the voter list) and could therefore be used by an academic researcher to link the identity of specific data subjects to the attributes in the released dataset. The background knowledge held by the attacker (the researcher) could be as follows: "I know that the data subject in the voter list with a specific triplet {ZIP code, gender, full date of birth} is unique. A record exists in the released dataset with that triplet". It was empirically observed[30] that the vast majority (more than 80%) of data subjects in the public register used in this research experiment was univocally associated to a specific triplet, which made the identification possible. Accordingly, the data were not properly anonymised in this case.

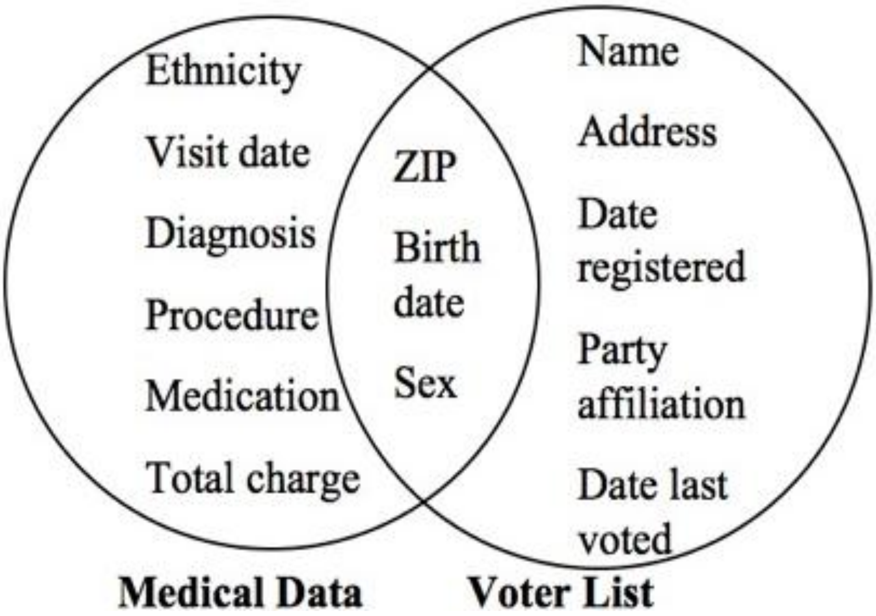


Figure A1. Re-identification by data linking

In order to reduce the effectiveness of similar linking attacks, it has been argued that controllers should first inspect the dataset and group those attributes that might reasonably be used by an attacker for linking the released table with another auxiliary source; each group should include at least *k* identical combinations of generalized attributes (i.e., it should represent an equivalence class of attributes). Datasets should then be released only after being partitioned into such homogeneous groups. The attributes selected for generalization are known in literature as quasi-identifiers, since their knowledge in clear would entail the immediate identification of data subjects.

Many identification experiments have shown the weakness of poorly engineered k-anonymised tables. This might happen, for instance, because the other attributes in an equivalence class are identical (as it happens for the equivalence class of Spanish data subjects in the example of Table A2) or their distribution is very unbalanced with a high prevalence of a specific attribute, or else because the number of records in an equivalence class is very low, allowing in both cases likelihood inference, or because no significant "semantic" difference exists between the attributes in clear of the equivalence classes (e.g., the quantitative measure of such attributes might be actually different but numerically very close, or they might belong to a range of semantically similar attributes, e.g. same range of

[30] L. Sweeney. Weaving Technology and Policy Together to Maintain Confidentiality. Journal of Law, Medicine &Ethics, 25, nos. 2&3 (1997): 98-110

credit risk, or same family of pathologies), so that the dataset is still capable of leaking a big amount of information on data subjects for linking attacks[31]. An important point to make here is that whenever data are sparse (for instance there are few occurrences of a specific property in a geographic area), and a first aggregation is not capable of grouping data with a sufficient number of occurrences of different properties (for instance still a low number of occurrences of few properties can be located in a geographic area), further attribute aggregation is necessary in order to reach targeted anonymisation.

l-diversity

Building upon these observations, variants of k-anonymity have been proposed over the years, and some engineering criteria for enhancing the practice of anonymisation by generalization have been developed, aimed at reducing the risks of linking attacks. They are based on probabilistic properties of datasets. Specifically, a further constraint is added, namely that each attribute in an equivalence class occurs at least *l* times, so that an attacker is always left with significant uncertainty on attributes even in the presence of background knowledge on a specific data subject. This is equivalent to saying that a dataset (or partition) should possess a minimum number of occurrences of a selected property: this trick might mitigate the risk of re-identification. This is the goal of the l-diversity anonymisation practice. An example of this practice is provided in tables A4 (the original data) and A5 (the result of the processing). As evident, by properly engineering the location ID and the ages of the individuals in table A4, generalization of attributes results into a substantial increase of the uncertainty as to the actual attributes of every data subject in the survey. For instance, even if the attacker knows that a data subject belongs to the first class of equivalence, he cannot further ascertain if a person possesses property X, Y or Z, since at least one record exists in that class (and in any other equivalence class) exhibiting such properties.

Serial number	Location ID	Age	Property
1	111	38	X
2	122	39	X
3	122	31	Y
4	111	33	Y
5	231	60	Z
6	231	65	X
7	233	57	Y
8	233	59	Y
9	111	41	Z
10	111	47	Z
11	122	46	Z
12	122	45	Z

Table A4. A table with individuals grouped by location, age and three properties X, Y and Z

[31] It has to be highlighted that correlations can also be established once data records have been grouped by attributes. When the data controller knows the types of correlations he wants to verify, he can select the attributes that are most relevant. For instance, PEW survey results are not subject to fine grained inference attacks and are still very useful to find correlations between demographics and interests (http://www.pewinternet.org/Reports/2013/Anonymity-online.aspx)

Serial number	Location ID	Age	Property
1	11*	<50	X
4	11*	<50	Y
9	11*	<50	Z
10	11*	<50	Z
5	23*	>50	Z
6	23*	>50	X
7	23*	>50	Y
8	23*	>50	Y
2	12*	<50	X
3	12*	<50	Y
11	12*	<50	Z
12	12*	<50	Z

Table A5. An example of 1-diverse version of table A4

t-closeness:

The specific case of attributes within a partition which are unevenly distributed or belong to a small range of values or semantic meanings is addressed by the approach known as t-closeness. This is a further improvement of anonymisation by generalization and consists in the practice of arranging data in order to achieve equivalence classes that mirror the initial distribution of attributes in the original dataset as much as possible. A two-step procedure is used for this purpose basically as follows. Table A6 is the original database including records in clear of data subjects, grouped by location, age, salary and two families of semantically similar properties, respectively (X1,X2, X3) and (Y1, Y2, Y3) (e.g. similar credit risk classes, similar diseases). First the table is *l-diversified* with *l*=1 (table A7), by grouping records into semantically similar equivalence classes and poor targeted anonymisation; then it is processed in order to obtain t-closeness (table A8) and higher variability within each partition. In fact, with the second step, each equivalence class includes records from both families of properties. It is worth noticing that location ID and age have different granularities in the various steps of the process: this means that each attribute might require different generalization criteria in order to obtain targeted anonymisation and this requires, in turn, specific engineering and appropriate computational burden by data controllers.

Serial number	Location ID	Age	Salary	Property
1	1127	29	30K	X1
2	1112	22	32K	X2
3	1128	27	35K	X3
4	1215	43	50K	X2
5	1219	52	120K	Y1
6	1216	47	60K	Y2
7	1115	30	55K	Y2
8	1123	36	100K	Y3
9	1117	32	110K	X3

Table A6. A table with individuals grouped by location, age, salaries and two families of properties

Serial number	Location ID	Age	Salary	Property
1	11**	2*	30K	X1
2	11**	2*	32K	X2
3	11**	2*	35K	X3
4	121*	>40	50K	X2
5	121*	>40	120K	Y1
6	121*	>40	60K	Y2
7	11**	3*	55K	Y2
8	11**	3*	100K	Y3
9	11**	3*	110K	X3

Table A7. An *l*-diverse version of table A6

Serial number	Location ID	Age	Salary	Property
1	112*	<40	30K	X1
3	112*	<40	35K	X3
8	112*	<40	100K	Y3
4	121*	>40	50K	X2
5	121*	>40	120K	Y1
6	121*	>40	60K	Y2
2	111*	<40	32K	X2
7	111*	<40	55K	Y2
9	111*	<40	110K	X3

Table A8. A *t*-close version of table A6

It has to be clearly stated that the goal of generalizing the attributes of data subjects in such educated ways can sometimes be achieved only for a small number of records, and not for all of them. Good practices should make sure that each equivalence class contains multiple individuals and that no inference attack remains possible. In any case, this approach requires in-depth assessment of the available data by data controllers along with the combinatorial evaluation of various alternatives (for instance, different range amplitudes, different location or age granularity etc.). In other words, anonymisation by generalization cannot be the outcome of a rough first attempt by data controllers to substitute analytical values of attributes in a record by ranges, as more specific quantitative approaches are needed - such as evaluating the entropy of attributes within each partition, or measuring the distance between the original attribute distributions and the distribution in each equivalence class.